UNCLASSIFIED

Information Science and Technology Seminar Speaker Series

Vladimir Braverman
Johns Hopkins University

Beating CountSketch for Heavy Hitters in Insertion Streams

Thursday, January 21, 2016

1:00 - 2:00 PM

TA-3, Bldg. 1690, Room 102 (CNLS Conference Room)

Abstract: Given a stream $p_1, \ldots, p_m$ of items from a universe $\mathcal{U}$, which, without loss of generality we identify with the set of integers $\{1, 2, \ldots, n\}$, we consider the problem of returning all ℓ_2-heavy hitters, i.e., those items j for which $f_j \geq \eps \sqrt{F_2}$, where f_j is the number of occurrences of item j in the stream, and $F_2 = \sum_{i \in [n]} f_i^2$. Such a guarantee is considerably stronger than the ℓ_1-guarantee, which finds those j for which $f_j \geq \eps m$. In 2002, Charikar, Chen, and Farach-Colton suggested the CountSketch data structure, which finds all such j using $\Theta (\log^2 n)$ bits of space (for constant $\eps > 0$). The only known lower bound is $\Omega(\log n)$ bits of space, which comes from the need to specify the identities of the items found.

In this paper we show it is possible to achieve $O(\log n \log \log n)$ bits of space for this problem. Our techniques, based on Gaussian processes, lead to a number of other new results for data streams, including:
(1) The first algorithm for estimating F_2 simultaneously at all points in a stream using only $O(\log n\log\log n)$ bits of space, improving a natural union bound and the algorithm of Huang, Tai, and Yi (2014).
(2) A way to estimate the ℓ_{∞} norm of a stream up to additive error $\eps \sqrt{F_2}$ with $O(\log n\loglog n)$ bits of space, resolving Open Question 3 from the IITK 2006 list for insertion only streams.

This is a joint work with Stephen R. Chestnut, Nikita Ivkin, and David P. Woodruff. The manuscript is available on http://arxiv.org/abs/1511.00661

Biography: Vladimir Braverman is an Assistant Professor with the Department of Computer Science at the Johns Hopkins University. His main research interests are streaming algorithms and their applications. Braverman obtained his B.Sc. and M.Sc. degrees from Ben-Gurion University of the Negev, Israel, and his Ph.D. from UCLA in 2011. Prior to attending UCLA, Braverman has led a research team at HyperRoll, a startup company that has been acquired by Oracle in 2009. Braverman has received the Edward K. Rice Outstanding Doctoral Student Award, the Google Outstanding Graduate Student Research Award, Outstanding Ph.D. Graduate Award of the Computer Science Department, UCLA and the Google Faculty Award. His research has been supported by the NSF and DARPA.

For more information contact the technical host Curt Canada, cvc@lanl.gov, 665-7453.

Hosted by the Information Science and Technology Institute (ISTI)

Los Alamos NATIONAL LABORATORY EST. 1943

ISTI INFORMATION SCIENCE & TECHNOLOGY INSTITUTE